Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

   Answer:

   - Month, Season and Weathersit have more impact on dependent variable cnt(Demand of Bike)
   - Demand of bike sharing, i.e. cnt, start growing at start of the year and is maximum during the middle of year in month of July and start declining towards the end of the year
   - Fall season have more demand of bike sharing as cnt is more than rest of the season
   - Clear weather have more demand of Bike sharing
   - Holiday, Weekday and Working Day do not have much impact on Demand of Bike

2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)

   Answer:
   Drop first is use to drop an extra column that is created during dummy variable creation. We perform encoding and decide on the value for the column without the column also.
   Ex: In the dataset season column contains 4 distinct values – spring, summer, fall and winter. So we can encode it as below with dropping the first column

spring	summer	fall	winter
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1

   Even if we drop the first column, we still should be able to understand that for first case when summer, fall and winter have 0 then it will be spring.

summer	fall	winter
0	**0**	**0**
1	0	0
0	1	0
0	0	1

   So we are are able to encode same thing and also remove one extra column from our data after using drop_first=True.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

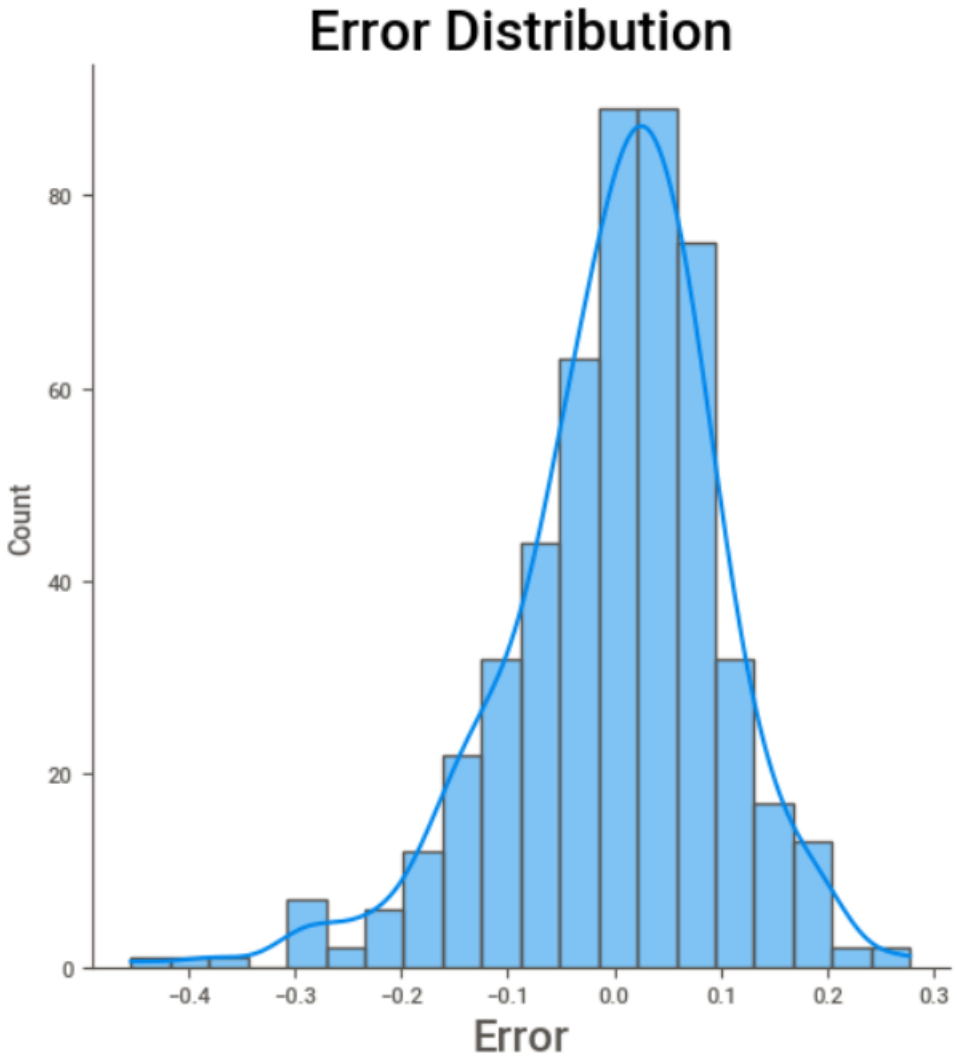
   Answer:
   From the pair plot among the numerical variables, **atemp and temp** have highest correlation with the target variable

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

   Answer:

To validate the assumptions of Linear Regression model we performed Residual analysis and on the final model that we created. Below are the steps performed.

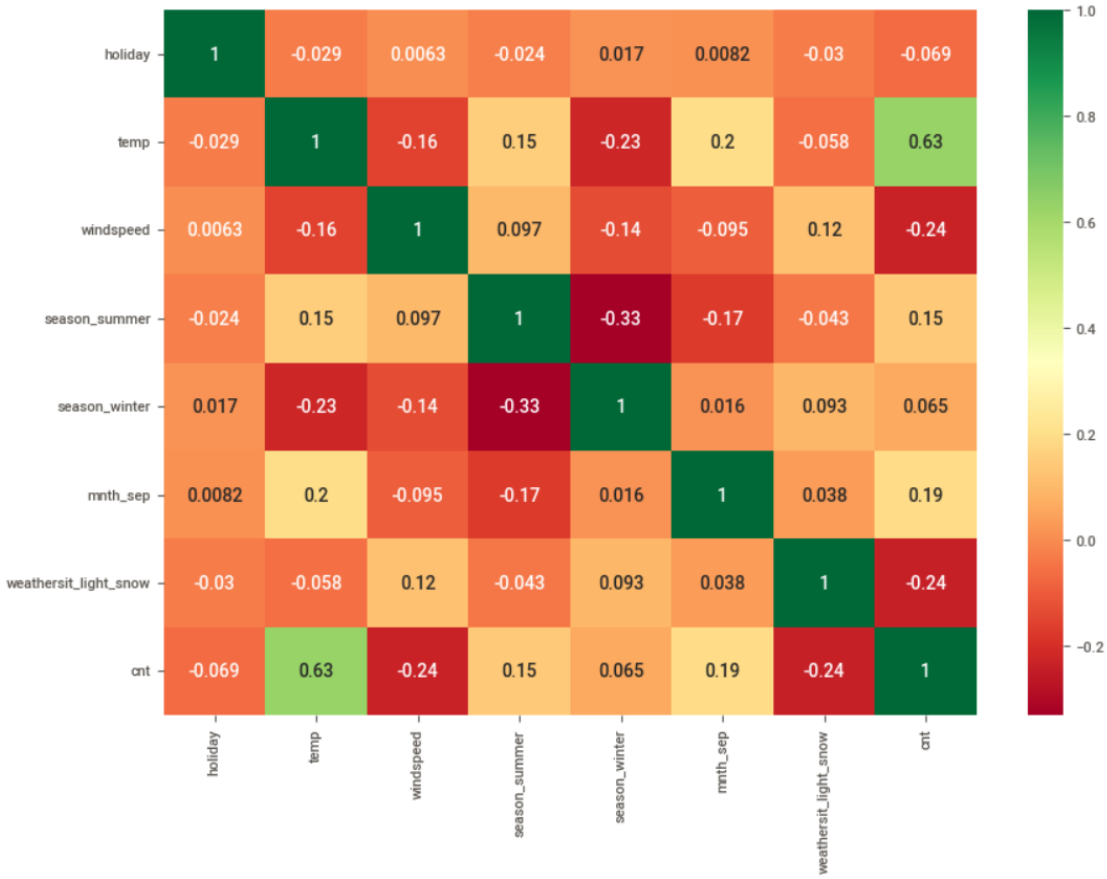
1. Get the predicted value of Y using the final model on training set
2. Get the residual. Residual(Error) = y_train(actual) – y_train_pred(predicted y values from model)
3. Plot distribution of Residual using displot.
4. The plotted graph should **have a normal distribution with mean at 0**. Same could be seen. Below you can see the plotted Error graph



5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Based on final model and correlation created for them, top 3 features contributing significantly explaining the demand are :

1. Temp
2. Month_sep
3. Season_summer



General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Answer:

Linear Regression is a statistical model which analyses the linear relationship between a dependent variable and independent variables. There are basically two types of linear regression –

a. Simple Linear Regression
b. Multiple Linear Regression

In Simple Linear Regression, we make prediction for one dependent variable with respect to one independent variable. Below is the formula for SLR.

$Y = c + mX$

where
Y = dependent variable
X = independent variable
c = intercept
m = slope

Below are some of the assumptions for simple linear regression
Assumptions SLR:

a. There should be linear relationship between X and Y.

b. Error term should be normally distributed with mean 0. X and Y is not needed to be normally distributed

In Multiple Linear Regression, we make prediction for one dependent variable with respect to more than one (multiple) independent variables. Below is the formula:

Y = c + m1X1 + m2X2 …… mnXn

In extension to assumptions of SLR, in MLR we need to consider few more aspects as below:

a. Overfitting
b. Multicollinearity
c. Feature Selection

2. Explain the Anscombe's quartet in detail. (3 marks)

Answer:

Anscombe's quartet consists of four datasets which is shows same statistical behaviour. It was constructed in 1973 by statistician Francis Anscombe. He used the to demonstrate both the importance of graphing data when analyzing it, and the effect of outliers and other influencial observations on statistical properties.

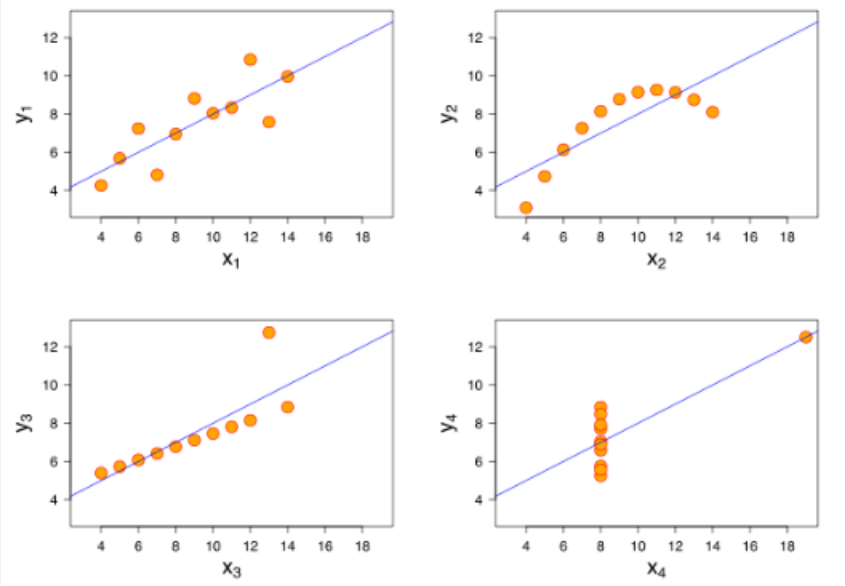
Below is the four datasets properties. As we can see the statistical property – mean, variance, correlation etc are same for all four datasets

Data [edit]

For all four datasets:

Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x : s_x^2	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y : s_y^2	4.125	±0.003
Correlation between x and y	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression : R^2	0.67	to 2 decimal places

Below is the figure when they were plotted. As we can observed, when plotted, they show different pattern



All four sets are identical when examined using simple summary statistics, but vary considerably when graphed

Below is the four datasets used in Anscombe quartet

Anscombe's quartet

I		II		III		IV	
x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

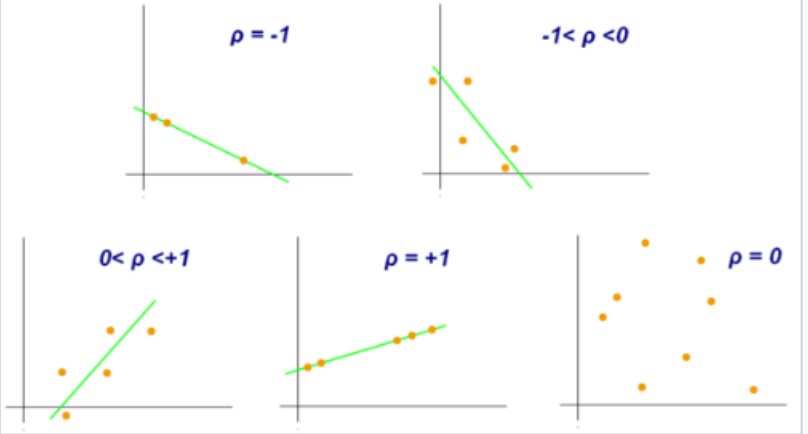
3. What is Pearson's R? (3 marks)

Answer:

Pearson's R is measure of linear correlation between two sets of data. It is ratio between the covariance of two variables and the product of their standard deviation.

If the variables tend to go up/down together then correlation is positive otherwise when

variables tend to go/down in opposite to each other then correlation is negative.
The correlation always varies between -1 and 1. The value near to 1 or -1 means the variables are highly correlated and near to 0 means the variables are not correlated.
Below is example of scatter diagram of different values of correlation.



Examples of scatter diagrams with different values of correlation coefficient (ρ)

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

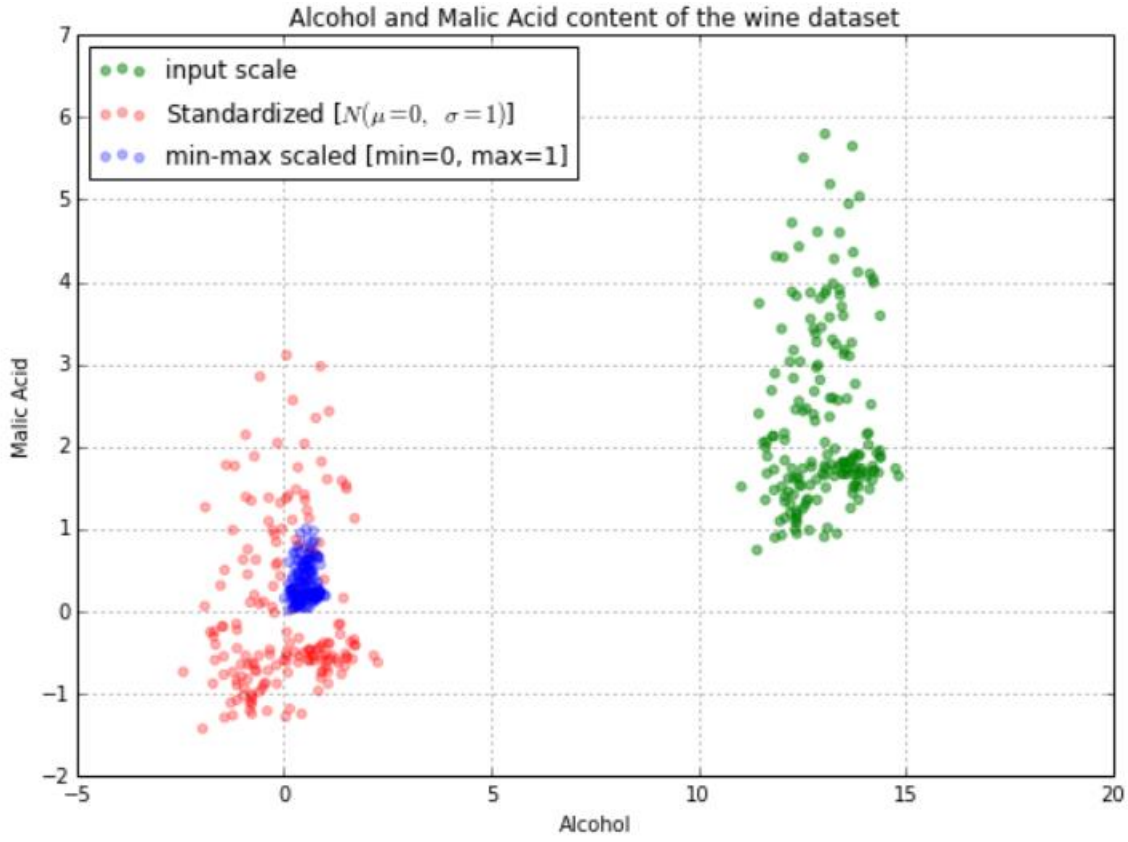
Answer:

Scaling is a way to bring all numerical features to same scale. When features are on same scale, it helps to evaluate the model better.
There are two types of scaling – normalized and standardized scaling.

Normalized scaling – This is also known as Min Max Scaling.
The formula used to scale is (X – Xmin)/(Xmax – Xmin). Using the formula, we always get minimum value as 0 and maximum as 1 all the features on which normalized scaling is performed.

Standardized scaling – This scaling method just brings the features into same scale. Formula used for this is (X – mean)/standard deviation.

Below graph explains two types of scaling the difference when they are plotted.



5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Answer:

When there is perfect correlation between independent variables, then VIF can increase and become infinite. So when one independent variable shows similar increase or decrease wrt to other independent variable.

Normally VIF < 5 is considered for a good feature. In case on infinite VIF, we should either drop one of the independent variable or can modify it like taking ratio, or applying some transformation so the correlation can be reduced.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

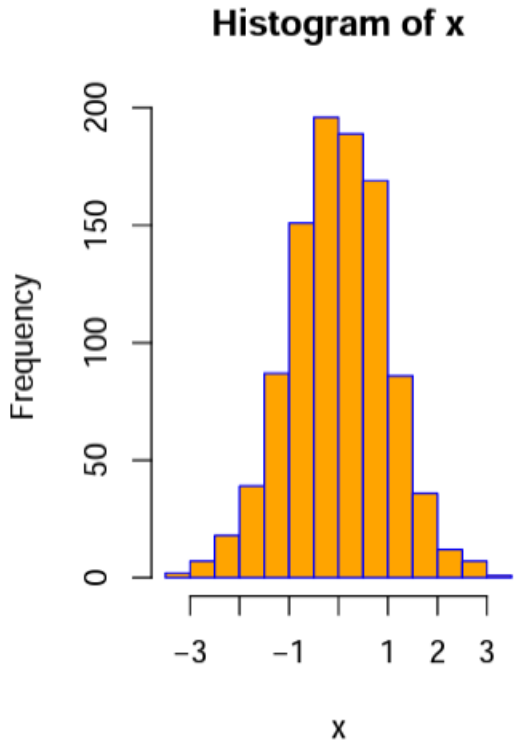
Q-Q plot is a plot when quantiles of two variables are plotted. It provides a summary of distribution of variables.

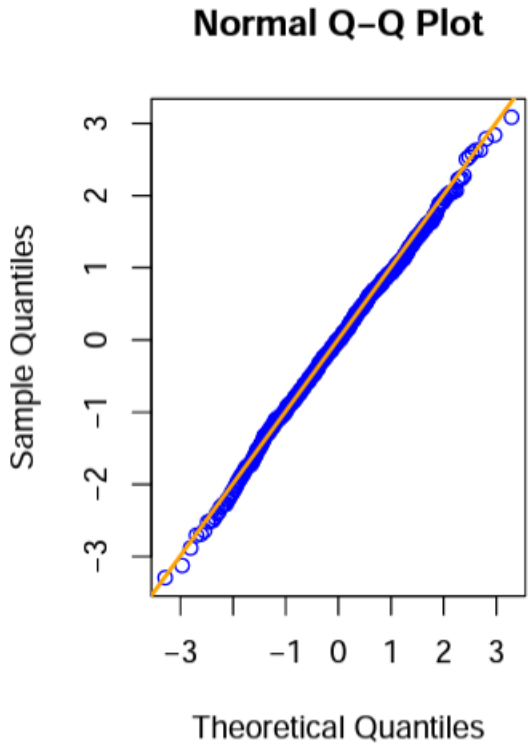
Use of Q-Q plot:

Q-Q show if two variables comes from same distribution or not. Quantile for first dataset is plotted along x-axis and second dataset along y-axis.

Many datasets comprised of normal distribution. So the normal distribution is the base distribution. The quantile plotted along x-axis is also know as Theoretical Quantiles and quantile of sample plotted along y-axis is known as Sample Quantiles.

Below we can see one of the examples of Q-Q plot.





Importance of Q-Q plot:
By looking at Q-Q plot we can understand if distribution in two dataset is similar or not.
If distribution is similar then point on Q-Q plot will lie exactly on y=x point. If distribution is linearly related then the point will lie on straight line but it will not be at y=x.